ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ, АНАЛИЗ ДАННЫХ, МАШИННОЕ ОБУЧЕНИЕ

УДК 519.85 ISSN 1995-5499
DOI: https://doi.org/10.17308/sait.2021.2/3509
Поступила в редакцию 13.03.2021
Подписана в печать 19.07.2021

КЛАСТЕРИЗАЦИЯ СОСТОЯНИЙ ПАЦИЕНТОВ ДЛЯ МОДЕЛИ НАЗНАЧЕНИЯ СХЕМ ЛЕЧЕНИЯ АТЕРОСКЛЕРОЗА

© 2021 М. В. Демченко, И. Л. Каширина✉, М. А. Фирюлина

Воронежский государственный университет,
Университетская пл., 1, 394018 Воронеж, Российская Федерация

Аннотация. В статье предлагается подход к реализации начального этапа решения задачи поиска и назначения оптимальных стратегий лечения пациентов с помощью моделей обучения с подкреплением, состоящего в выделении основных групп состояний пациентов с диагностированным атеросклерозом с использованием кластерного анализа. В качестве исходного набора данных была использована выборка MIMIC-III, содержащая значения клинических, лабораторных, гемодинамических и др. показателей пациентов. Основным методом кластерного анализа в данной работе был выбран метод k-medoids, при этом качество кластеризации оценивалось с помощью силуэтного анализа. Предварительным этапом кластеризации являлось понижение размерности с помощью метода главных компонент (PCA), а визуализация результатов производилась с помощью метода t-SNE. При этом важным этапом данного исследования являлось вычисление оценки тяжести состояния пациента для каждого из выявленных кластеров состояний. Полученные оценки используются для вычисления вознаграждений в модели назначения оптимальных схем лечения с помощью методов обучения с подкреплением, при этом набор полученных кластеров определяет набор состояний окружения. Таким образом, результаты кластеризации позволяют выявить основные закономерности в исходном наборе данных, а также позволяют сформировать основные составляющие модели обучения с подкреплением для назначения оптимальных схем лечения атеросклероза.
Ключевые слова: MIMIC-III, машинное обучение, кластеризация, k-medoids, понижение размерности, PCA, t-SNE, атеросклероз.

ВВЕДЕНИЕ

Исследование и разработка наиболее эффективных методов диагностики и лечения атеросклероза является важнейшей задачей современной медицины. Актуальность всестороннего исследования данного заболевания обусловлена тем, что оно начинает развиваться в раннем возрасте и является причиной поражения сосудов, существенно повышая риск возникновения сердечно-сосудистых заболеваний, представляющих собой угрозу здоровью и жизни человека.

Несмотря на то, что существуют (и ежегодно обновляются) международные клинические рекомендации по лечению атеросклероза [1], по-прежнему отсутствуют эффективные инструменты для персонифицированной поддержки принятия решений в реальном времени, опирающиеся на динамику изменения состояния пациента. С целью создания подобного инструментария предлагается использовать основанный на данных

✉ Каширина Ирина Леонидовна
e-mail: kash.irina@mail.ru

подход для определения стратегий лечения атеросклероза с использованием алгоритмов глубокого обучения с подкреплением.

Модели обучения с подкреплением основываются на вычислительном подходе к обучению, при котором целью *агента* является максимизация суммарного *вознаграждения*, которое он получает во время взаимодействия с окружением, как правило, сложным и неопределенным (рис. 1). Данный подход описывает марковский процесс [2], модель которого включает несколько основных составляющих.

1. Состояния — выделенные группы схожих состояний здоровья пациентов.

2. Действия — медицинские предписания и процедуры, выполняемые согласно плану лечения пациентов.

3. Вознаграждения — численная оценка влияния лечения на состояние пациента.

4. Вероятности изменения состояний — матрица вероятностей перехода пациента из одного состояния в другое при условии применения какого-либо из действий, т. е. вероятность изменения состояния здоровья пациента при условии проведения медицинских процедур или приема назначенных препаратов.

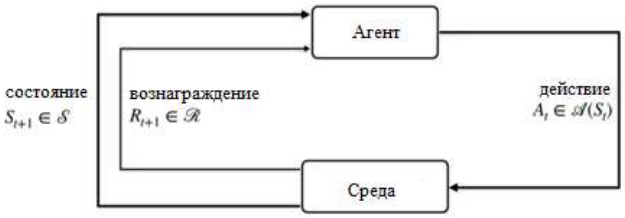

Рис. 1. Модель обучения с подкреплением [Fig. 1. Reinforcement learning model]

В настоящий момент такой подход является весьма актуальным. В частности, в работах [3, 4] продемонстрированы впечатляющие результаты, полученные при использовании алгоритмов обучения с подкреплением для выбора оптимальных схем лечения диабета, а в работе [5] модель обучения с подкреплением строится для назначения лечения при сепсисе. Однако в данных работах вознаграждения зависят от единственной характеристики — уровня глюкозы при диабете и от результатов выживаемости пациентов при сепсисе. В случае атеросклероза такой единственный маркер отсутствует. При этом многообещающим подходом является использование нетривиальных функций вознаграждения, связанных с общей оценкой текущего состояния пациента.

Как отмечено выше, начальным этапом создания модели обучения с подкреплением является выделение набора состояний пациентов, таких что процесс лечения в динамике характеризуется изменением этих состояний.

Эффективным способом реализации задачи выделения множества состояний модели является кластерный анализ [5], где каждый кластер отражает соответствующее состояние здоровья пациентов.

При этом выбор и реализация алгоритма кластеризации имеет важное значение. Необходимо, чтобы полученные кластеры были плотными и хорошо отделимыми, то есть каждый кластер включал набор очень схожих состояний пациентов. При этом число кластеров должно быть достаточно большим, так как в процессе лечения атеросклероза применяется несколько различных групп препаратов и важно иметь возможность отследить, как динамически меняется состояние конкретного пациента под воздействием той или иной комбинации лекарственных средств.

Оптимальное разбиение показателей здоровья пациентов на кластеры позволит сформировать конечное множество состояний модели обучения с подкреплением, при этом для каждого из кластеров вычисляется оценка степени тяжести соответствующего состояния пациента. Вычисленные оценки в дальнейшем используются в качестве вознаграждений модели обучения с подкреплением.

Таким образом, данное исследование посвящено формированию модели обучения с подкреплением для назначения оптимальных схем лечения, а именно, этапу определения основных групп состояний здоровья пациентов.

Данная работа является продолжением решения задачи, посвященной разработке интеллектуальных алгоритмов ранней диагностики и лечения атеросклероза, поставленной в 2014 году Воронежским областным кардиологическим диспансером [6, 7], который предоставил деперсонифицированные

данные о 522 пациентах. В работах [8–10] отражены основные результаты проведенного ранее исследования.

К сожалению, имеющаяся в распоряжении авторов российская база не содержит результатов динамического наблюдения за пациентами, необходимых для построения моделей обучения с подкреплением, поэтому в исследовании использовалась открытая база данных MIMIC-III, содержащая такую информацию [11, 12].

Из данной базы были выбраны пациенты, для которых измерялись гемодинамические, лабораторные, антропометрические, социально-демографические и клинические характеристики, аналогичные тем, что присутствовали в российской базе исследований. Это сделано для того, чтобы, построенную модель впоследствии можно было масштабировать на отечественные наборы данных.

В результате исходный набор данных о пациентах с атеросклерозом, извлеченный из базы данных MIMIC-III, содержит 10670 записей измерений о госпитализациях 331 пациента и включает 79 признаков (табл. 1). Запись набора данных представляет собой показатели пациента, измеряемые с определенной периодичностью.

Схемы лечения атеросклероза, представленные в базе MIMIC-III, соответствуют международным рекомендациям. Основная цель исследования — показать потенциальную применимость алгоритмов обучения с подкреплением для персонифицированной поддержки принятия решений по назначению лечения атеросклероза, опирающиеся на динамику изменения состояния пациента.

МАТЕРИАЛЫ И МЕТОДЫ

Одним из этапов, связанных с качественным проведением кластерного анализа, является предварительное понижение размерности, так как исходное количество признаков (79) может являться избыточным.

Данный подход во многих случаях позволяет улучшить качество и скорость классификации, а также провести визуализацию полученных результатов с помощью представления исходных данных в двумерном или трехмерном пространствах.

Предварительное понижение размерности исходного набора данных было проведено с использованием метода главных компонент [13], который позволяет выделить набор наиболее значимых признаков для кластери-

Таблица 1. Основные признаки набора данных
[Table 1. Dataset features]

Гемодинамические	АД (артериальное давление), ДАД (диастолическое АД), САД (систолическое АД), ЦВД (центральное венозное давление), сердечный индекс, SVRI, частота сердечных сокращений (ЧСС), частота дыхания, SpO2 %, сердечный ритм (нормальный синусовый, синусовая аритмия, мерцательная аритмия и др.)
Лабораторные	Глюкоза, холестерин, креатинин, гемоглобин, магний, тромбоциты, лейкоциты, эритроциты, анионная разница, бикарбонаты, хлориды, натрий, калий, гематокрит, протромбиновое время (INR), средний объем эритроцитов (MCV), ширина распределения эритроцитов (RDW), активированное частичное тромбопластиновое время (АЧВТ), PH, pCO2, pO2 и др.
Антропометрические	Вес, пол, возраст
Социально-демографические	Семейное положение
Клинические	Тип госпитализации (плановый, срочный, экстренный)

зации таким образом, чтобы сохранить большую часть информации исходной выборки.

Алгоритм понижения размерности с помощью метода главных компонент

Шаг 1. Стандартизовать исходный набор данных.

Шаг 2. Вычислить матрицу ковариации $\Sigma = \left(\sigma_{jk}\right)$ по формуле (1).

$$\sigma_{jk} = \frac{1}{n}\sum_{i=1}^{n}\left(x_j^{(i)} - \mu_j\right)\left(x_k^{(i)} - \mu_k\right), \qquad (1)$$

где μ_j и μ_k — средние значения признаков x_j и μ_k $j = \overline{1,N}$, $k = \overline{1,N}$, N — число признаков (размерность) исходного пространства, n — число записей.

Шаг 3. Вычислить собственные векторы и собственные значения матрицы Σ.

Шаг 4. Отсортировать собственные вектора в порядке убывания собственных значений.

Шаг 5. Выбрать N' собственных векторов, соответствующих N' максимальным собственным значениям, где N' — размерность нового пространства.

Объясняемая дисперсия для признака x_j вычисляется по формуле (2).

$$Explained \quad variance = \frac{\lambda_j}{\sum_{k=1}^{N}\lambda_k}, \qquad (2)$$

где λ_j — собственное значение, соответствующее признаку x_j.

Шаг 6. Вычислить матрицу проекции W с помощью выбранных N' собственных значений.

Шаг 7. Сформировать из исходного набора X данных новый набор данных X' размерности N' с помощью матрицы проекции W по формуле (3).

$$X' = XW. \qquad (3)$$

При выборе алгоритма кластеризации для решения данной задачи одной из определяющих характеристик являлась интерпретируемость получаемых центров кластеров, что важно для задачи оценки состояния пациента. Одним из эффективных и интерпретируемых методов кластеризации является метод k-медоидов [14]. Входным параметром данного алгоритма является число предполагаемых кластеров. Преимуществом данного метода является алгоритм выделения центроидов (медоидов), в результате которого центры кластеров являются существующими точками исходного набора данных (в отличие от широко известного алгоритма k-средних). Данный подход, в частности, применялся в исследованиях [15, 16].

Начальное положение кластеров предлагается выбирать по алгоритму, который инициализирует центроиды таким образом, чтобы они были удаленными друг от друга, что, как правило, приводит к лучшим результатам, чем случайная инициализация. Аналогичный подход используется в методе k-means++ [17]

Алгоритм кластеризации k-medoids

Шаг 1. Случайным образом выбрать из набора исходных данных первую центральную точку C_1.

Шаг 2. Вычислить расстояние от всех точек набора данных (исключая уже назначенные центроидами) до всех имеющихся центроидов C_j. Для каждой точки x_i найти расстояние d_i до ближайшего к ней центроида

$$d_i = \min_{1 \le j \le m}\left\|x_i - C_j\right\|^2, \qquad (4)$$

где m — число уже зафиксированных центроидов, $1 \le m < k$.

Шаг 3. Для каждой точки x_i, рассчитать вероятность быть выбранной в качестве $m+1$ центроида по формуле: $p_i = \frac{d_i}{\sum_{j=1}^{r} d_j}$, где r — общее число точек, не назначенных к данному шагу центроидами.

Шаг 4. В качестве очередного центроида случайным образом выбрать точку x_i в соответствии с полученным распределением вероятностей.

Шаг 5. Повторять шаги 2–4 до тех пор, пока не будут найдены k центроидов.

Шаг 6. Для каждой точки набора данных, вычислить расстояние между точкой и всеми найденными центроидами. Точка будет определена к кластеру, соответствующему ближайшему к ней центроиду.

Шаг 7. Произвести пересчет центроидов в кластерах. Выбрать в качестве нового набора центроидов k точек исходных данных $(C_1, C_2, ..., C_k)$, минимизирующих функцию потерь (5) для каждого кластера S_i.

$$C_i = \underset{C \in S_i}{argmin} \sum_{x \in S_i} \|x - C\|^2. \quad (5)$$

Шаг 8. Повторять шаги 6, 7 до тех пор, пока центроиды не перестанут меняться.

Качество полученной кластеризации оценивалось с помощью силуэтного анализа. Коэффициент силуэта является метрикой, отражающей качество кластеризации с учетом расстояния между кластерами и между объектами внутри кластеров.

$$Silhouette = \frac{b - a}{\max(b, a)}, \quad (6)$$

где b — среднее расстояние между точками из различных кластеров, a — среднее расстояние между точками внутри кластеров. Данная метрика изменяется в пределах от –1 до 1, и чем она ближе к единице, тем качественнее результат кластеризации (то есть тем более плотные и отделимые кластеры удалось сформировать).

Одним из наиболее востребованных и современных методов нелинейного понижения размерности является метод t-SNE, с помощью которого была произведена визуализация результатов полученной кластеризации [18].

3. РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Предварительным этапом проведения кластерного анализа являлось понижение размерности с помощью метода главных компонент. Данный этап позволил существенно повысить качество и ускорить процесс кластеризации за счет фильтрации признаков, связанных с информационным шумом, и выделением части признаков, которые несут наибольший объем информации.

Оценка количества главных компонент производилась с использованием вычисления объясняемой дисперсии. График накопленной объясняемой дисперсии приведен на рис. 2. Исходя из полученных значений, в качестве главных компонент кластерного анализа было выбрано 14 признаков, объясняющих 83 % дисперсии.

Качество кластеризации, проведенной после этапа понижения размерности, с помощью метода k-medoids, оценивалось с помощью силуэтного анализа, и эта же метрика использовалась для выбора оптимального числа кластеров. Метод силуэта был выбран потому, что, как уже отмечалось, необходимо было выделить достаточно плотные и отделимые кластеры состояний пациентов с малым внутрикластерным и одновременно достаточно большим межкластерным расстоянием. Популярный на практике метод локтя (каменистой осыпи) больше ориентирован на уменьшение внутрикластерных расстояний, а метод силуэта одновременно учитывает и внутрикластерные и межкластерные расстояния.

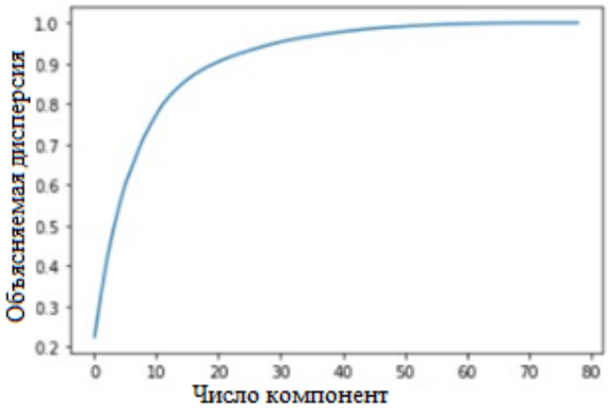

Рис. 2. График объясняемой дисперсии
[Fig. 2. Cumulative explained variance]

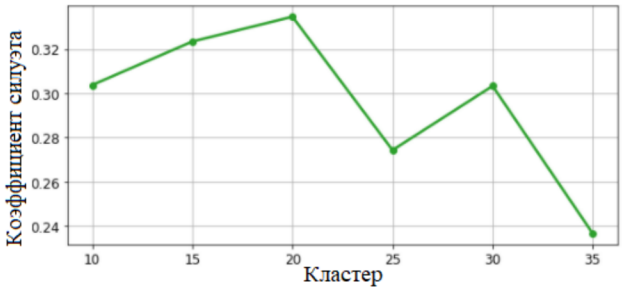

Рис. 3. Силуэтный анализ
[Fig. 3. Silhouette analysis]

Рис. 3 иллюстрирует график значений коэффициента силуэта в зависимости от выбранного числа кластеров, который достигает максимума (0.33) при выбранных 20 кластерах.

Данное количество кластеров вполне соответствует цели исследования, так как является достаточно большим для того, чтобы описать наиболее характерные различные состояния пациентов, проходящих лечение от атеросклероза, динамически меняющиеся под воздействием той или иной комбинации лекарственных средств и воздействий.

Визуализация полученных результатов кластеризации с помощью метода t-SNE приведена на рис. 4. Данный график иллюстрирует 20 кластеров состояний пациентов, при этом степень тяжести состояния здоровья пациента является низкой в кластерах 1–9, средней в кластерах 10–14 и высокой в кластерах 15–20. Данный график явно демонстрирует, что по результатам кластеризации наиболее многочисленной является группа кластеров, связанная с тяжелыми состояниями здоровья, группы средней тяжести располагаются на графике по центру на близком расстоянии друг от друга, при этом наиболее многочисленные кластеры, связанные с низкой степенью тяжести состояния, расположены удаленно от кластеров, связанных с высоким риском для здоровья.

График медоидов приведен на рис. 5, исходя из которого можно сделать вывод, что центральными точками кластеров являются объекты с характеристиками, отличающимися между собой.

Табл. 2 представляет собой фрагмент сводной таблицы анализа полученных кластеров.

Кластер 1 содержит минимальный процент отклонений от нормы большинства показателей (глюкозы, гемоглобина, эритроцитов, PH, понижений SVRI), а также низкий процент смертности в кластере.

Кластеры 3 и 5 в среднем соответствуют низкому проценту отклонений от нормы показателей, при этом в кластере 3 зафиксирован минимальный процент отклонений от нормы холестерина, в кластере 5 — сердечного индекса.

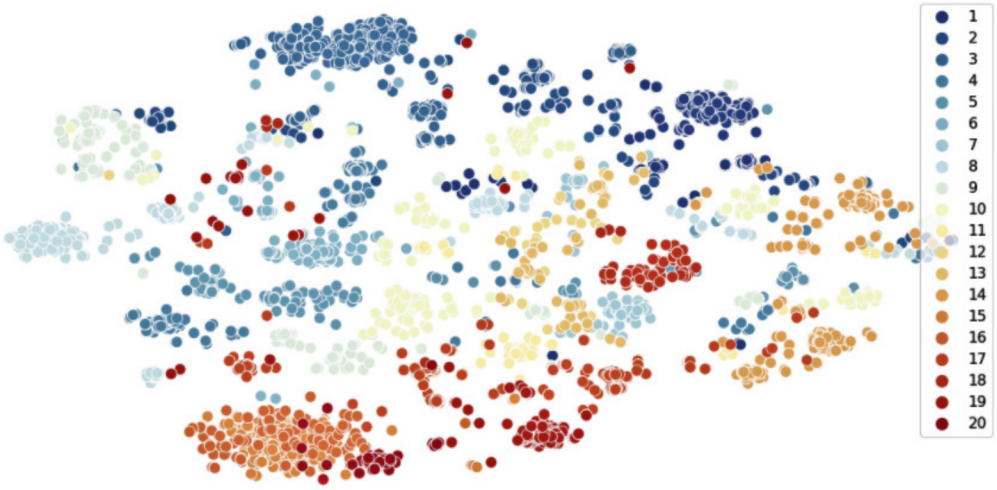

Рис. 4. Визуализация результатов кластеризации
[Fig. 4. Clustering visualization]

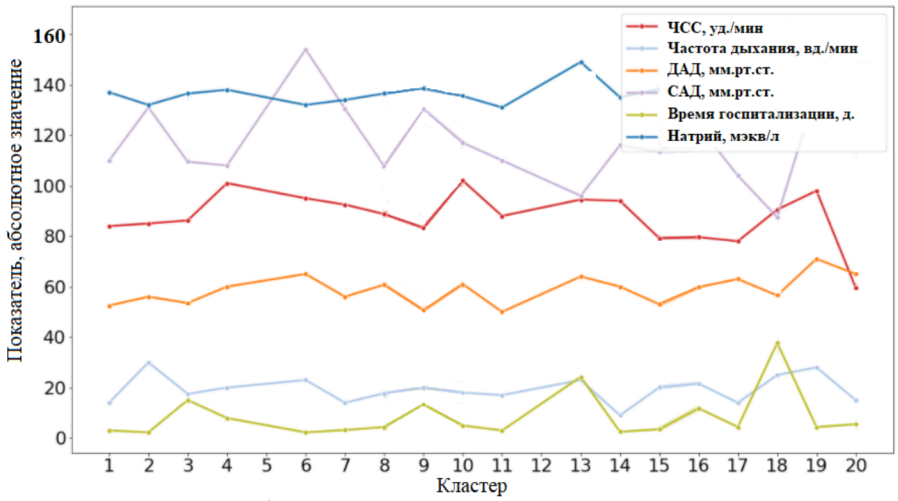

Рис. 5. График центральных элементов кластеров
[Fig. 5. Medoids of clusters]

Таблица 2. Анализ полученных кластеров
[Table 2. Cluster analysis]

Кластер / Показатель	1	3	5	8	10	12	15	18	20
Холестерин [вне нормы], %	9	*0.9*	4.5	12.7	14.7	5.8	16.3	13.1	**35.6**
Глюкоза [вне нормы], %	*29.1*	35.6	37.8	34.4	35	33	41.1	40	**53.1**
Гемоглобин [вне нормы], %	*36.4*	45.1	49.1	43	43	46.9	51.5	**55.5**	44.3
Эритроциты [вне нормы],%	*32.9*	44	49.6	41.5	40.5	47.1	51.1	**55.5**	42.7
РН [вне нормы], %	*9.6*	16	16.7	10.5	10.1	11.5	**17.6**	13	16.5
АЧТВ [вне нормы], %	16.6	23	14.7	17.0	16.3	28.4	19.7	*13.9*	**46.6**
SVRI [ниже нормы], %	*32.6*	39.9	36.5	48.2	53.6	56.3	**67.8**	52.1	42.1
SVRI [выше нормы], %	25.7	30.7	29.8	17.4	13.4	17.2	16.4	*12.2*	**63.4**
Сердечный индекс [ниже нормы], %	40.9	38.4	32.6	37.6	*25.2*	50.1	37.9	47.7	**62.1**
Сердечный индекс [выше нормы], %	5.7	5.9	*4.9*	7.8	**13.9**	5	8.2	4.9	13.6
АД [ниже нормы], %	11.4	10.5	11.6	12.8	11.6	**18.4**	13.1	16.0	*8.4*
АД [выше нормы], %	0.6	3.6	1.7	0.5	1.6	*0.4*	1.6	0.5	**12.3**
ЦВД [ниже нормы], %	0.0	0.1	0.8	0.1	0.2	*0*	*0*	*0*	**2.6**
ЦВД [выше нормы], %	1.6	1.9	1.2	1.9	2.1	**5.8**	3.9	0.3	*0*
ЧСС [ниже нормы], %	0.1	0.7	0.4	0.2	*0*	0.2	0.7	1.2	**9.4**
ЧСС [выше нормы], %	0.2	5.0	2.9	0.4	**6.5**	1.2	0.4	*0*	2.9
Доля умерших, %	0.8	1.8	3.2	0.6	1.6	*0*	*0*	**16.7**	*0*

В кластере 8 не зафиксировано минимальных или максимальных значений отклонений от нормы показателей. Пониженных значений ЦВД и смертности не было зафиксировано в кластерах 12 и 15, отклонений ЧСС ниже и выше нормы не наблюдалось в кластерах 10 и 12, соответственно. Однако в кластере 8 наблюдалось частое повышение ЧСС и сердечного индекса, в кластере 12 — пониженное АД и повышенное ЦВД, в кластере 15 — пониженное SVRI и отклонения от нормы РН. В кластере 18 зафиксирован наиболее высокий процент смертности, частые отклонения от нормы гемоглобина и эритроцитов. Кластер 20 ассоциирован с наиболее частыми отклонениями от нормы большинства показателей: холестерина, глюкозы, АЧТВ, SVRI, АД (выше нормы), сердечного индекса, ЦВД, ЧСС (ниже нормы).

Таким образом, в результате проведенной кластеризации было получено 20 кластеров, каждому из которых была назначена оценка, отражающая степень тяжести состояния пациента, состояние которого соответствует данному кластеру и принимающая значения в пределах [0,1].

Для получения этой оценки:

– для всех показателей пациента был вычислен процент отклонений этого показателя от нормы в данном кластере;

– был вычислен процент смертности в данном кластере;

– в качестве итоговой оценки степени тяжести состояния пациента было выбрано среднее арифметическое процента отклонений от нормы показателей и смертности, значение нормируется в диапазоне [0,1]. При построении оценки учитывались проведенные ранее исследования маркеров атеросклероза [6–10].

График, отражающий шкалу оценок каждого из кластеров, приведен на рис. 6.

Данные оценки (взятые с отрицательным знаком) представляют собой вознаграждения модели обучения с подкреплением при переходе между состояниями (полученными кластерами).

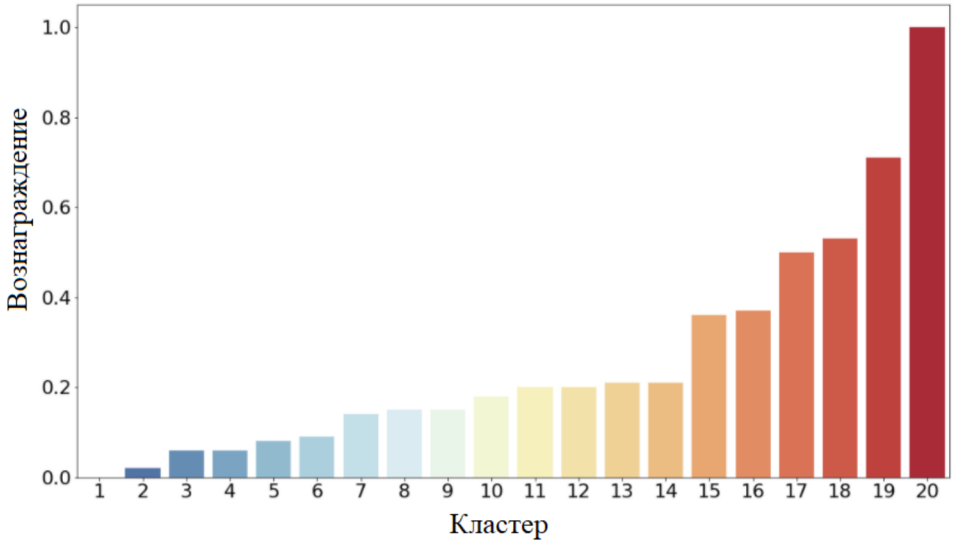

Рис. 6. График вознаграждений в модели обучения с подкреплением
[Fig. 6. Reinforcement learning model rewards]

ЗАКЛЮЧЕНИЕ

В результате работы были реализованы этапы понижения размерности, кластерного анализа и визуализации полученных кластеров для набора состояний пациентов с диагностированным атеросклерозом с помощью таких методов машинного обучения, как k-medoids, PCA и t-SNE.

Результаты кластеризации позволяют выявить основные закономерности в состояниях, соответствующих такому заболеванию, как атеросклероз, на основе группы гемодинамических, лабораторных и др. характеристик пациентов.

Проведенная кластеризация позволяет описать пространство возможных различных состояний пациентов и разработать нетривиальные функции вознаграждения при динамическом изменении состояния пациента в процессе лечения.

Полученные результаты будут использованы при дальнейшей разработке модели обучения с подкреплением, целью которой является назначение и корректировка персонифицированных стратегий лечения данного заболевания в реальном времени.

БЛАГОДАРНОСТИ

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 20-37-90029 Аспиранты

КОНФЛИКТ ИНТЕРЕСОВ

Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

СПИСОК ЛИТЕРАТУРЫ

1. *Gerhard-Herman M. D. et al.* A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines // Circulation. Ovid Technologies (Wolters Kluwer Health). – 2016. – Vol. 135, No 12. – P. 686–725.

2. Sutton, R. Reinforcement learning / R. Sutton, A. Barto. – Cambridge : The MIT Press, 2018. – 526 p.

3. *Yom-Tov, E.* Encouraging physical activity in patients with diabetes: Intervention using a reinforcement learning system / E. Yom-Tov, G. Feraru, M. Kozdoba, S. Mannor, M. Tennenholtz, I. Hochberg // Journal of medical Internet research. – 2017. – Vol. 19, No 10. – P. e338. DOI: 10.2196/jmir.7994

4. *Noori, A.* Glucose level control using Temporal Difference methods / A. Noori, M. A. Sadrnia, M. B. N. Sistani // 2017 Iranian Conference on Electrical Engineering (ICEE). – 2017. – P. 895–900. DOI: 10.1109/IranianCEE.2017.7985166

5. *Komorowski, M.* The artificial intelligence clinician learns optimal treatment strategies

for sepsis in intensive care / M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, A. A. Faisal // Nature Medicine. – 2018. – Vol. 24, No 11. – P. 1716-1720. DOI: 10.1038/s41591-018-0213-5

6. *Хохлов, Р. А.* Предикторы атеросклеротического поражения артерий конечностей по данным кардиоангиологического скрининга взрослого населения / Р. А. Хохлов, А. Э. Гайдашев, Н. М. Ахмеджанов // Рациональная фармакотерапия в кардиологии. – 2015. – Т. 11, №5. – С. 470–476. DOI: 10.20996/1819-6446-2015-11-5-470-476

7. *Хохлов, Р. А.* Использование многоканальной объемной сфигмографии для кардиоангиологического скрининга взрослого населения / Р. А. Хохлов, Н. И. Остроушко, А. Э. Гайдашев, Д. В. Кирсанов, Н. М. Ахмеджанов // Рациональная фармакотерапия в кардиологии. – 2015. – Т. 11, № 4. – С. 371–379. DOI: 10.20996/1819-6446-2015-11-4-371-379

8. *Demchenko, M. V.* The development of the atherosclerosis diagnostic models under conditions of unbalanced classes / M. V. Demchenko, I. L. Kashirina // J. Phys.: Conf. Ser. – 2020. – Vol. 1479, 012026. DOI: 10.1088/1742-6596/1479/1/012026

9. *Demchenko, M.* The use of machine learning methods to the automated atherosclerosis diagnostic and treatment system development/ M. Demchenko, I. Kashirina// CEUR Workshop Proceedings. – 2020. – 2790. – P. 233–245.

10. *Львович, Я. Е.* Использование методов машинного обучения для исследования маркеров атеросклероза магистральных артерий / Я. Е. Львович, И. Л. Каширина, М. В. Демченко // Информационные технологии. – 2020. – Т. 26, № 5. – С. 46–55. DOI: 10.17587/it.26.46-55

11. *Johnson, A.* MIMIC-III, a freely accessible critical care database / A. Johnson, T. Pollard, L. Shen // Sci Data. – 2016. – Vol. 3, 160035. DOI: 10.1038/sdata.2016.35

12. *Dai, Z.* Analysis of adult disease characteristics and mortality on MIMIC-III / Z. Dai, S. Liu, J. Wu, M. Li, J. Liu, K. Li // PLOS ONE. – 2020. – Vol. 15, e0232176. DOI: 10.1371/journal.pone.0232176

13. *Pearson, K.* On lines and planes of closest fit to systems of points in space / K. Pearson // Philosophical Magazine. – 1901. – Vol. 2. – P. 559–572. DOI: 10.1080/14786440109462720

14. *Reynolds, A. P.* The Application of K-Medoids and PAM to the Clustering of Rules / A. P. Reynolds, G. Richards, V. J. Rayward-Smith // Intelligent Data Engineering and Automated Learning – IDEAL 2004. – Springer, Berlin, Heidelberg, 2004. – Vol. 3177. – P. 173–178. DOI: 10.1007/978-3-540-28651-6_25

15. *Yan, J.* Applying Machine Learning Algorithms to Segment High-Cost Patient Populations / J. Yan, K. A. Linn, B. W. Powers, et al. // Journal of General Internal Medicine. – 2019. – Vol. 34, № 2. – P. 211–217. DOI: 10.1007/s11606-018-4760-8

16. *Zhang, Q.* A New and Efficient K-Medoid Algorithm for Spatial Clustering / Q. Zhang, I. Couloigner // Lecture Notes in Computer Science. – 2005. – Vol. 3482. – P. 181-189. DOI: 10.1007/11424857_20

17. *Arthur, D.* K-Means++: The Advantages of Careful Seeding / D. Arthur, S. Vassilviskii // Proc. of the Annu. ACM-SIAM Symp. on Discrete Algorithms. – 2007. – Vol. 8. – P. 1027–1035. DOI: 10.1145/1283383.1283494

18. *Van der Maaten, L. J. P.* Visualizing High-Dimensional Data Using t-SNE / L. J. P. van der Maaten, G.E. Hinton // Journal of Machine Learning Research. – 2008. – Vol. 9. – P. 2579–2605.

Демченко Мария Владиславовна — аспирант факультета ПММ Воронежского государственного университета.
E-mail: masha-vrn@yandex.ru
ORCID: https://orcid.org/0000-0002-6439-8957

Каширина Ирина Леонидовна — д-р техн. наук, профессор кафедры математических методов исследования операций факультета ПММ Воронежского государственного университета.
E-mail: kash.irina@mail.ru
ORCID: https://orcid.org/0000-0002-8664-9817

Фирюлина Мария Андреевна – аспирант факультета ПММ Воронежского государственного университета.
E-mail: mashafiryulina@mail.ru
ORCID: https://orcid.org/0000-0003-3468-5514

УДК 519.85 ISSN 1995-5499
DOI: https://doi.org/10.17308/sait.2021.2/3509
Received 13.03.2021
Accepted 19.07.2021

CLUSTER ANALYSIS OF PATIENTS' STATES PERFORMED IN ORDER TO DEVELOP TREATMENT STRATEGIES FOR PATIENTS WITH ATHEROSCLEROSIS

© 2021 M. V. Demchenko, I. L. Kashirina✉, M. A. Firyulina

Voronezh State University
1, Universitetskaya Square, 394018 Voronezh, Russian Federation

Annotation. The article describes an approach to the implementation of the initial stage of solving the problem of finding and prescribing optimal treatment strategies using reinforcement learning. The approach involves the identification of the main groups of conditions of patients with diagnosed atherosclerosis by means of cluster analysis. The MIMIC-III database containing the clinical, laboratory, hemodynamic, and other data of patients was used as the initial data set. The main cluster analysis method used in the study was the k-medoids algorithm. The quality of clustering was assessed by means of silhouette analysis. At the preliminary stage of clustering, we reduced the dimensionality using principal component analysis (PCA). The results were visualized using the t-SNE method. An important part of the study was the calculation of the severity of the patients' conditions for each of the identified clusters. The resulting estimates were then used to calculate the rewards in the model for assigning optimal treatment plans by means of reinforcement learning. The set of obtained clusters determines the set of the environment states. Thus, the clustering results allowed us to identify the main patterns in the initial dataset and to obtain the main components of the reinforcement learning model for prescribing optimal treatment plans for atherosclerosis.
Keywords: MIMIC-III, machine learning, clustering, k-medoids, dimension reduction, PCA, t-SNE, atherosclerosis.

✉ Kashirina Irina L.
e-mail: kash.irina@mail.ru

CONFLICT OF INTEREST

The authors declare the absence of obvious and potential conflicts of interest related to the publication of this article.

REFERENCES

1. *Gerhard-Herman M. D. et al.* (2016) A Report of the American College of Cardiology. American Heart Association Task Force on Clinical Practice Guidelines . Circulation. Ovid Technologies. Vol. 135, No 12. P. 686–725.

2. *Sutton R.* (1992) Reinforcement learning. Boston : Kluwer Academic Publishers.

3. *Yom-Tov E., Feraru G., Kozdoba M., Mannor S., Tennenholtz M. and Hochberg I.* (2017) Encouraging Physical Activity in Patients With Diabetes: Intervention Using a Reinforcement Learning System. Journal of Medical Internet Research. 19(10). e338. Available from: doi: 10.2196/jmir.7994

4. *Noori A., Sadrnia M. and Sistani M.* (2017) Glucose level control using Temporal Difference methods. 2017 Iranian Conference on Electrical Engineering (ICEE). P. 895–900. Available from: doi:10.1109/IranianCEE.2017.7985166

5. *Komorowski M., Celi L., Badawi O., Gordon A. and Faisal A.* (2018) The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nature Medicine. 24(11). P. 1716–1720. Available from: doi: 10.1038/s41591-018-0213-5

6. *Khokhlov R., Gaydashev A. and Akhmedzhanov N.* (2015) Predictors of atherosclerotic lesions of limb arteries according to cardioangiological screening of the adult population. Rational Pharmacotherapy in Cardiology. 11(5). P. 470–476. (In Russian). Available from: doi: 10.20996/1819-6446-2015-11-5-470-476

7. *Khokhlov R., Ostroushko N., Gaydashev A., Kirsanov D. and Akhmedzhanov N.* (2015) Multi-channel volume sphygmography in cardioangiological screening of the adult population. Rational Pharmacotherapy in Cardiology. 11(4). P. 371–379. (In Russian). Available from: doi: 10.20996/1819-6446-2015-11-4-371-379

8. *Demchenko M. and Kashirina I.* (2020) The development of the atherosclerosis diagnostic models under conditions of unbalanced classes. Journal of Physics: Conference Series, 1479. 012026. Available from: doi: 10.1088/1742-6596/1479/1/012026

9. *Demchenko M. and Kashirina I.* (2020) The Use of Machine Learning Methods to the Automated Atherosclerosis Diagnostic and Treatment System Development. In: CEUR Workshop Proceedings. 2790. P. 233–245.

10. *Lvovich Y.* (2020) The Use of Machine Learning Methods to Study Markers of Atherosclerosis of the Great Arteries. Informacionnye tehnologii. 26(1). P. 46–55. Available from: doi: 10.17587/it.26.46-55

11. *Johnson A., Pollard T., Shen L., Lehman L., Feng M., Ghassemi M., Moody B., Szolovits P., Anthony Celi L. and Mark R.* (2016) MIMIC-III, a freely accessible critical care database. Scientific Data. 3(1). Available from: doi: 10.1038/sdata.2016.35

12. *Dai Z., Liu S., Wu J., Li M., Liu J. and Li K.* (2020) Analysis of adult disease characteristics and mortality on MIMIC-III. PLOS ONE. 15(4). e0232176. Available from: doi: 10.1371/journal.pone.0232176

13. *Pearson K.* (1901) LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 2(11). P. 559-572. Available from: doi: 10.1080/14786440109462720

14. *Reynolds A., Richards G. and Rayward-Smith V.* (2004). The Application of K-Medoids and PAM to the Clustering of Rules. Lecture Notes in Computer Science. P. 173–178. Available from: doi:10.1007/978-3-540-28651-6_25

15. *Yan J., Linn K., Powers B., Zhu J., Jain S., Kowalski J. and Navathe A.* (2018) Applying Machine Learning Algorithms to Segment High-Cost Patient Populations. Journal of General Internal Medicine. 34(2). P. 211–217. Available from: doi:10.1007/s11606-018-4760-8

16. *Zhang Q. and Couloigner I.* (2005) A New and Efficient K-Medoid Algorithm for Spatial Clustering. Computational Science and Its Applications – ICCSA 2005. P. 181–189. Available from: doi:10.1007/11424857_20

17. *Arthur D and Vassilvitskii S.* (2007) K-Means++: The Advantages of Careful Seeding. Proc. of the Annu. ACM-SIAM Symp. on Dis-

crete Algorithms. P. 1027–1035. Available from: doi: 10.1145/1283383.1283494

18. *Van der Maaten L. and Hinton G.* (2011) Visualizing non-metric similarities in multiple maps. Machine Learning. 87(1). P. 33–55.

Demchenko Maria V. — postgraduate student, Faculty of Applied Mathematics and Mechanics, Voronezh State University.
E-mail: masha-vrn@yandex.ru
ORCID: https://orcid.org/0000-0002-6439-8957

Kashirina Irina L. — DSc in Technical Sciences, Professor, Department of Mathematical Methods of Operations Research, Faculty of Applied Mathematics and Mechanics, Voronezh State University.
E-mail: kash.irina@mail.ru
ORCID: https://orcid.org/0000-0002-8664-9817

Firyulina Maria A. – postgraduate student, Faculty of Applied Mathematics and Mechanics, Voronezh State University
E-mail: mashafiryulina@mail.ru
ORCID: https://orcid.org/0000-0003-3468-5514